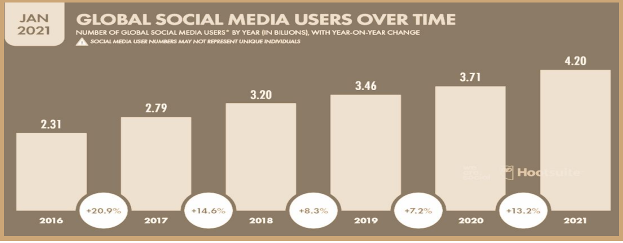
Toxic Comment Classification

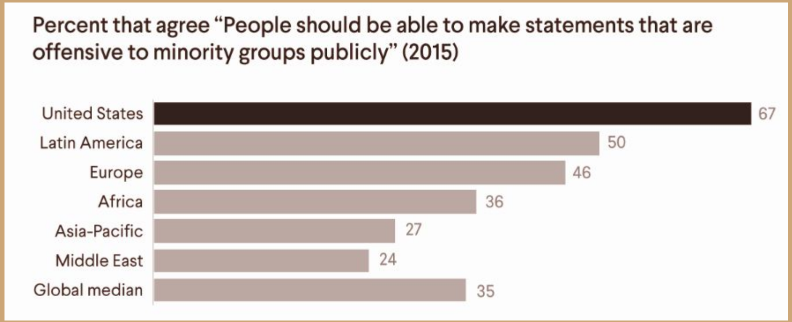
Helping online platforms identify toxicity, one comment at a time

Team Blue:
Amrita Ligga
Rizabek Zhumkenov
Scott Wais
Sudhanshu Rai
Sumeet Duddagi

Toxicity - a widespread problem

Toxic content growth





How Social Media Spurred Myanmar's Latest Violence

Everybody will end up losing if hate speech is left unchecked.

Effect on Online Platforms

A $150 billion lawsuit over genocide may force Facebook to confront its dark side

More than 1,200 families suing social media companies over kids' mental health

Twitter Tops The List Of Most Toxic Apps

Toxic online content harms moderators' mental health

Problem Statement

The internet is a place where people can post their views easily, resulting in challenges for online platforms to regulate what is being posted. These platforms might struggle to prevent online abuse or harassment of different communities. Not handling toxic comments can cause several problems for the platforms:

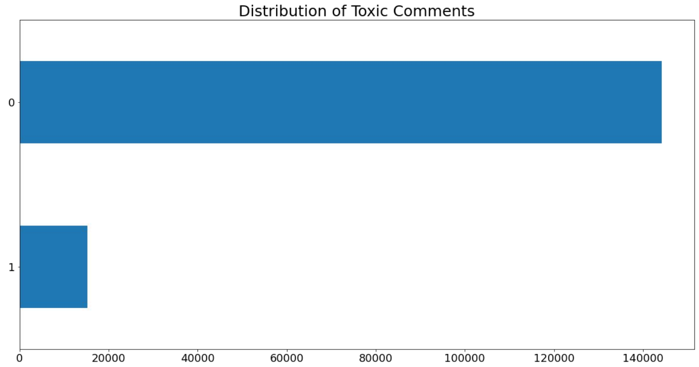
- Lawsuits
- Damage to reputation
- Reduced traffic
- Reduced Ad revenue

Using this dataset and text analytics techniques, we aim to build a model to identify toxic comments for such online platforms.

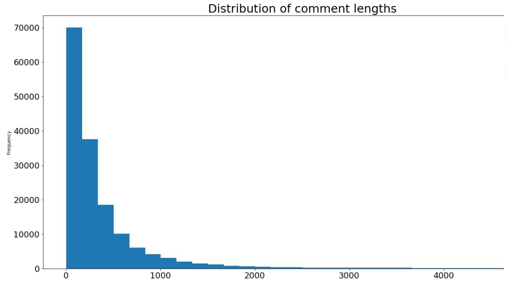
Data

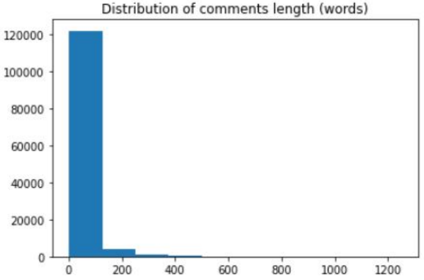
- 159,571 Wikipedia comments
- 3 Columns: id, comment_text, toxic label

	id	comment_text	toxic
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0
2	000113f07ec002fd	Hey man, I'm really not trying to edit war. It...	0
3	0001b41b1c6bb37e	"\nMore\nI can't make any real suggestions on ...	0
4	0001d958c54c6e35	You, sir, are my hero. Any chance you remember...	0



- Toxic comments: ~10%
- Comment Length Distribution: Right skewed with most comments between 0-500 characters or 0-100 words long





Cleaning and Data Preparation

Average Count Caps Column

- Many toxic comments use capital letters to emphasize strong emotions
- This column calculates the following:
- [total capital letters/total letters]
- This was done before any pre-processing

Word Replacements

- Word contractions: "mustn't" -> "must not"
- Regex Groupings using Textacy [URLs, emails, currency, emojis]
- Custom Regex Groupings [IP Addresses, Dates, Child]

Word Removals

- Non English Words
- Special characters like ~, *, +, =
- Sentence punctuations
- HTML characters like \n
- Stopwords

Final Preparation

- Spelling corrections using probability theory
- Tokenization
- Lemmatization and stemming
- Train test split (80:20)

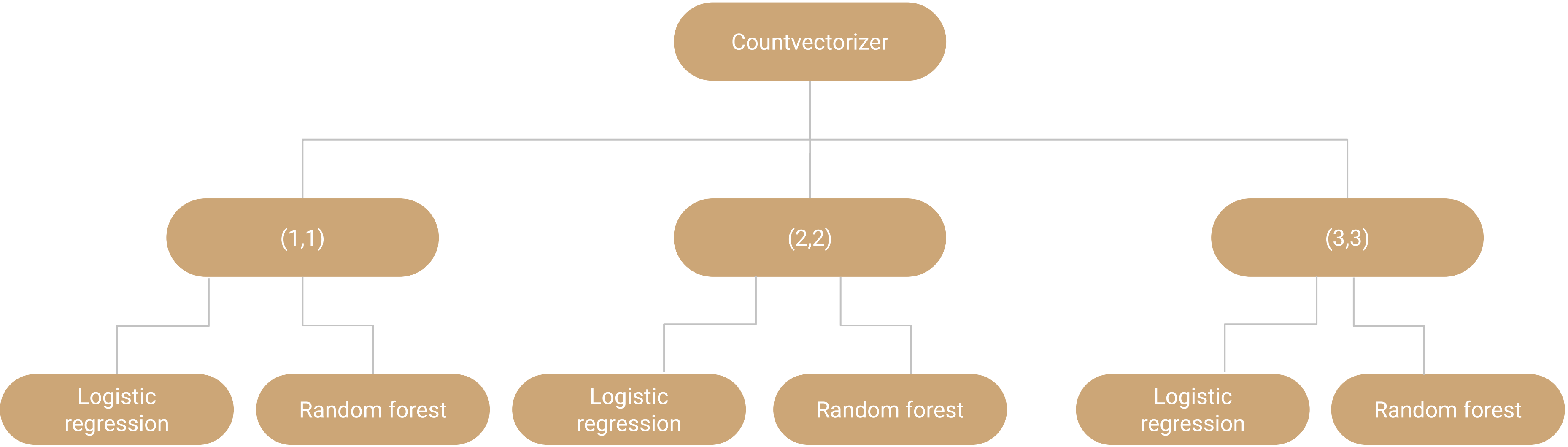
Data Cleaning Samples

Original	Cleaned(with stopwords)
You are wrong about everything \n\nbut for too stupid to notice. You are a lost cause. Your existence is a blemish.137.205.183.70	you are wrong about everything but for too stupid to notice you are a lost cause your existence is a blemish _ipaddress_
27 January 2010 (UTC)\n\nI have filed a complaint against Cshay for edit warring.\n22:18	_datetime_ utc i have filed a complaint against cshay for edit warring _time_
I only comunicate you. I can discute the article because it´s judging by Wikipedia moderators.	i only communicate you i can dispute the article because it s judging by wikipedia moderators

Created different versions of cleaned columns for example with and without stopword removal since some algorithms could work better with stopwords included:

id	comment_text	avg_count_caps	comment_cleaned	comment_tokenized	comment_cleaned_spell	comment_cleaned_no_stopwords	comment_cleaned_spell_no_stopwords	comment_cleaned_no_stopwords_lemm	comment_cleaned_spell_no_stopwords_lemm	comment_cleaned_lemm	comment_cleaned_spell_lemm	toxic
be631fee6a996d52	List of Moroccan Dutch people \n\nPeople shoul...	0.054422	list of moroccan dutch people people should ...	[List, of, Moroccan, Dutch, people, People, sh...	list of moroccan dutch people people should ha...	list moroccan dutch people people wikipedia pa...	list moroccan dutch people people wikipedia pa...	list moroccan dutch people people wikipedia pa...	list moroccan dutch people people wikipedia pa...	list of moroccan dutch people people should ha...	list of moroccan dutch people people should ha...	0
955b143f75755ac3	Not much better, be careful should someone rep...	0.012658	not much better be careful should someone rep...	[Not, much, better, be, careful, should, someo...	not much better be careful should someone repo...	much better careful someone report ban vandalism	much better careful someone report ban vandalism	much well careful someone report ban vandalism	much well careful someone report ban vandalism	not much good be careful should someone report...	not much good be careful should someone report...	0
e623261850dd2aaa	I blocked the pair of you because you are not ...	0.022857	i blocked the pair of you because you are not ...	[I, blocked, the, pair, of, you, because, you,...	i blocked the pair of you because you are not ...	blocked pair write encyclopaedia plenty places...	blocked pair write encyclopaedia plenty places...	block pair write encyclopaedia plenty place pl...	block pair write encyclopaedia plenty place pl...	i block the pair of you because you be not her...	i block the pair of you because you be not her...	0
068e0a8032a4cf5c	Thank You \n\nThanks for the link on the refer...	0.020833	thank you thanks for the link on the referen...	[Thank, You, Thanks, for, the, link, on, the, ...	thank you thanks for the link on the reference...	thank thanks link reference desks crux orthodo...	thank thanks link reference desks crux orthodo...	thank thanks link reference desk crux orthodox...	thank thanks link reference desk crux orthodox...	thank you thank for the link on the reference ...	thank you thank for the link on the reference ...	0
dd852a6da913a096	I know of about 50,000 Bible Scholars who disa...	0.034146	i know of about _number_ bible scholars who di...	[I, know, of, about, _NUMBER_, Bible, Scholars...	i know of about _number_ bible scholars who di...	know _number_ bible scholars disagree yet libe...	know _number_ bible scholars disagree yet libe...	know _number_ bible scholar disagree yet liber...	know _number_ bible scholar disagree yet liber...	i know of about _number_ bible scholar who dis...	i know of about _number_ bible scholar who dis...	0

Countvectorizer



- We used lemmatized comments with no stop words for countvectorization. We used the above model combinations to test on testing set using auc roc score, recall, and precision.
- The dataset was initially undersampled with 10,000 toxic and 10,000 non-toxic entries, and this was utilized as a training set and evaluated on test set.
- Next, the complete dataset was utilized as a training set and tested on test set.
- Finally, undersampled dataset performed comparatively better through logistic regression model trigrams with recall of 97.91%, precision of 10.21%, auc_roc score of 0.53 and f1 score of 0.18

TF-IDF Vectorizer

- Vectorized comments with unigrams, bigrams, and trigrams
  - Used comments that had been cleaned, spell checked, and lemmatized (stopwords removed)
  - Concatenated vectorized comments with avg_count_caps feature
- For each feature set, built three Logistic Regressions and Random Forests
  - Experimented with different probability cutoffs around the true toxic proportion [0.05,0.1,0.15]
  - Evaluated performance using precision, recall, and f1-score
- Best performance:

features	model	cutoff	precision	recall	f1
TF-IDF (unigrams)	**Logistic Regression**	**0.15**	0.423	0.66	0.515

- ROC AUC for this model: **0.891**

Logistic regression (Word2Vec)

- Vectorized comments using average word2vec embedding
  - Used comments that had been cleaned, spell checked, and lemmatized (stopwords removed)
- Built a Logistic Regression model for the feature set
  - Put probability cutoffs 0.5
  - Evaluated performance using precision, recall, and f1-score
- Best performance:
    - Roc_auc_score: 0.95
    - Recall: 0.56
    - Precision: 0.82
    - F1_score: 0.66

Deep learning model using pre-trained word embedding

Input data	Embeddings	Document embeddings	Flatten and Dense	Testing result
Training data set contains **127,656** words (lemmatized). We set the padding length **200** based on histogram below. Input matrix size is 127,656x200	We used ***glova*** embedding that has size **100**. Vocab size of training set is 138,467. Embedding matrix is 138,467 x 100 x 200	We defined that **138,467 x 100** parameters as **non-trainable.**	Flatten layer has 20,000 parameters (200 x 100). Dense = 1 Total **20,001 trainable** parameters.	roc_auc_score: **0.87** Recall: **0.52** Precision: **0.67** F1_score: **0.52**

Deep Learning (RNN and LSTM)

RNN	LSTM
● Number of Tokens: 10,000 ● Max Sequence Length: 200 ● Glove Embeddings (size:100) ● Total Parameters: 18,582,417 ● Number of Trained Parameters: 11,617 `precision_6: 0.0956 - recall_6: 1.0000 - auc_4: 0.5000`	● Number of Tokens: 10,000 ● Max Sequence Length: 200 ● Glove Embeddings (size:100) ● Total Parameters: 18,588,369 ● Number of Trained Parameters: 17,569 `precision_7: 0.0958 - recall_7: 1.0000 - auc_5: 0.5000`

Summary: The deep learning models did not perform well giving low precision of ~10%, due to data imbalance.

All Models

Model	ROC-AUC	Precision	Recall	F1-Score
Logistic Regression (Word2Vec embeddings)	**0.95**	**0.82**	**0.56**	**0.58**
DL pre-trained embedding glova	0.87	0.67	0.52	0.52
Logistic Regression (CountVectorizer)	0.53	0.10	0.97	0.18
Logistic Regression (TF-IDF)	0.89	0.42	0.66	0.52
RNN	0.5	0.0956	1.00	0.174
LSTM	0.5	0.0958	1.00	0.175

The NLP model could reduce the content moderation cost **by 76%**

# total comments	10,000
% of toxic comments	10%
# of toxic comment	1,000

	Simple Keywords searches	**Best NLP Model**
accuracy	78%	95%
recall	80%	56%
precision	29%	82%
FP comments to review	2000	114
TP comments to review	800	547
Total comments to review	2800	662
Cost per each review by human	$2.00	$2.00
Total cost	**$2,802.00**	**$663.78**

Comments
Simple keyword search is an ineffective method of identifying toxic comments. This method has high recall but leads to high rates of false positives.

The table compares cost of human moderation for 10,000 comments.

Limitations, Risk and Assumptions

- The case assumes that the company uses simple keyword search to moderate comments. We didn’t compare our model with existing solutions on the market
- Although the best NLP model reduces the number of comments to be reviewed by humans, the number of false negative comments will increase
- These false negatives carry a hidden cost because they could lead to bad user experiences on our platform. We are assuming that our increase in precision will outweigh this decrease in recall
- Even though our dataset was smaller than a real-world dataset, some of our models required a relatively longer time to run and we were limited by less processing power

Conclusion

- Toxicity is a major problem on online platforms and advanced NLP techniques can be used to reduce it with lesser human intervention
- After preprocessing, we tried various models from simple to complex to identify toxic comments
- The best model is Logistic Regression with Average Word2Vec embeddings, providing an F1 score of 0.58 and AUC of 0.95
- The approximate ROI is a **76%** reduction in moderation cost

Further Improvements (BERT)

If given more time we would do the following:

- Collect better data in consultation with experts
- We attempted implementing tiny-BERT but could have tried larger models given more processing power
- We could also explore hyper parameter tuning given more time